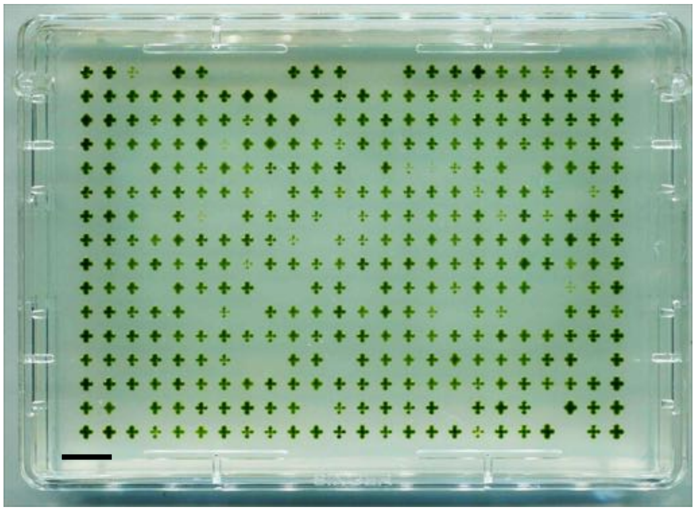
A



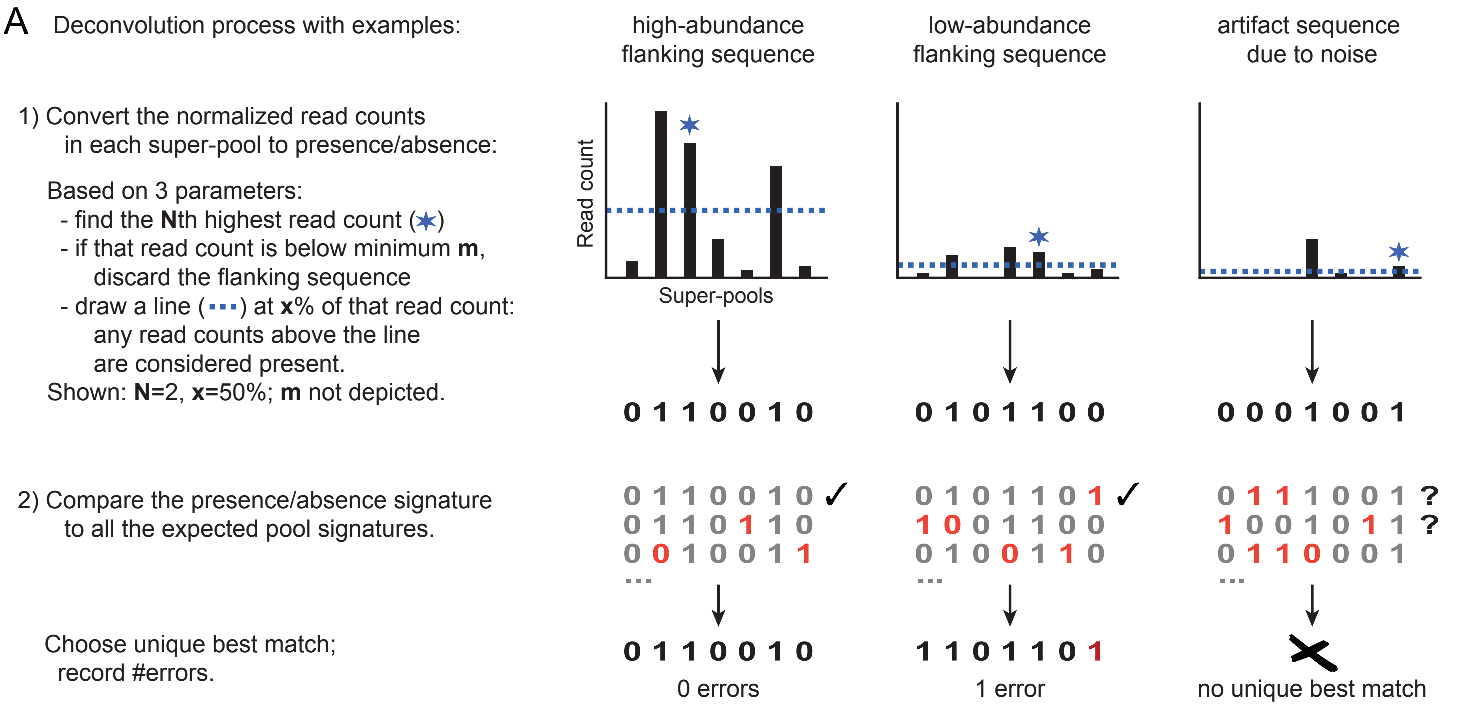
B

Colony loss	Number	Percent of colonies originally present
Whole-plate contamination and recovery	335	1.82 %
Partial contamination and treatment	25	0.14 %
Random loss	379	2.06 %
Total	739	4.01 %

Colony gain	Number	Percent of originally empty spots
Total	67	3.49 %

Supplemental Figure 1. Mutants can be robustly maintained as arrays of colonies on agar.

(A) A representative mutant library plate, imaged two weeks after it was generated. Bar = 1 cm.
(B) We quantified loss and gain of colonies over the course of 18 months of propagating the library.



A Deconvolution process with examples:

1) Convert the normalized read counts in each super-pool to presence/absence:

Based on 3 parameters:
- find the **N**th highest read count (✶)
- if that read count is below minimum **m**, discard the flanking sequence
- draw a line (▪▪▪) at **x**% of that read count: any read counts above the line are considered present.

Shown: **N**=2, **x**=50%; **m** not depicted.

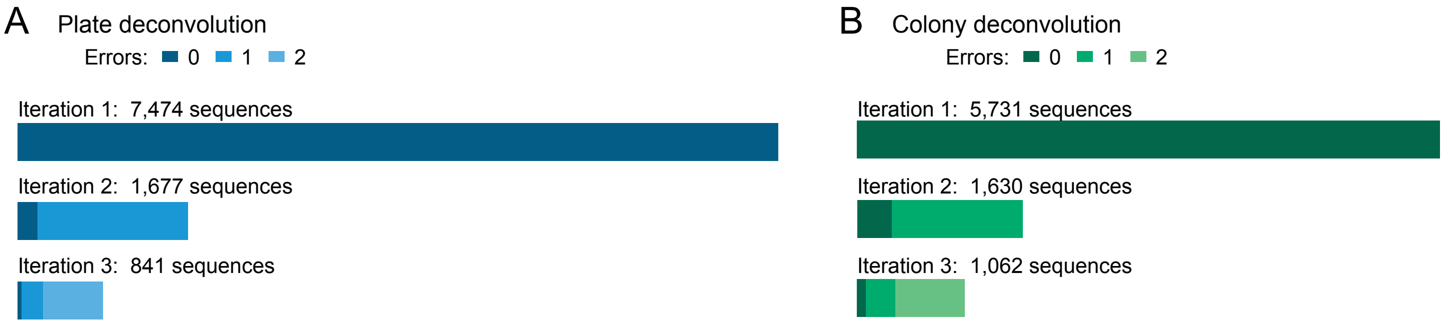
2) Compare the presence/absence signature to all the expected pool signatures.

Choose unique best match; record #errors.

B Optimized deconvolution parameters:	5’ plate data	3’ plate data	5’ colony data	3’ colony data
Iteration 1: optimized for quality, 0 errors	N=6, x=7%, m=20	N=6, x=10%, m=20	N=8, x=20%, m=20	N=8, x=20%, m=30
Iteration 2: intermediate, 0-1 errors	N=5, x=5%, m=20	N=5, x=7%, m=10	N=7, x=10%, m=30	N=7, x=10%, m=10
Iteration 3: optimized for quantity, 0-2 errors	N=4, x=20%, m=10	N=4, x=20%, m=10	N=6, x=10%, m=10	N=6, x=20%, m=10

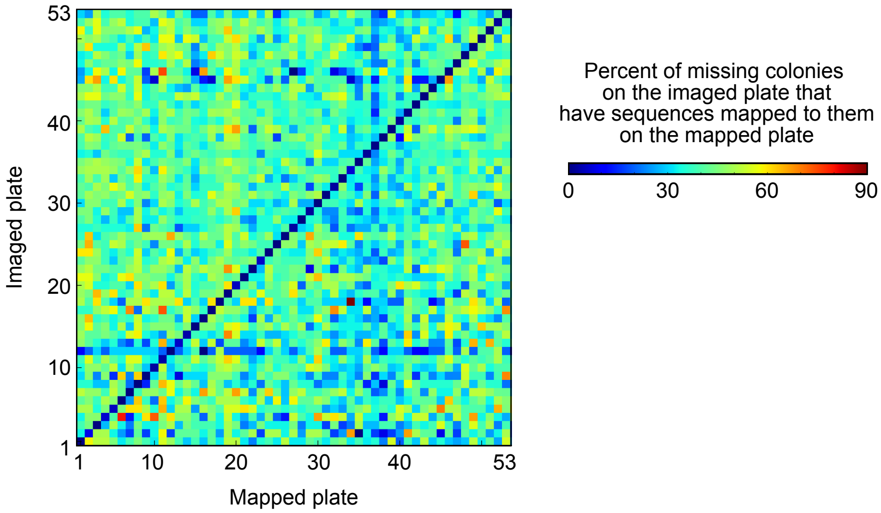
Supplemental Figure 2. Flanking sequences for each super-pool are computationally deconvolved into plate and colony coordinates.

(A) The deconvolution process is explained visually with examples.
(B) The optimized parameters used for three iterations of plate and colony deconvolution are shown (see Supplemental Methods).



Supplemental Figure 3. Details of flanking sequence deconvolution results separated into three iterations.

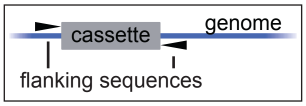
(A) The numbers of flanking sequences mapped to a plate during plate deconvolution iterations 1-3, with different numbers of errors.
(B) The same plot is shown for colony deconvolution.

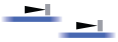


Supplemental Figure 4. Validation of deconvolution results.

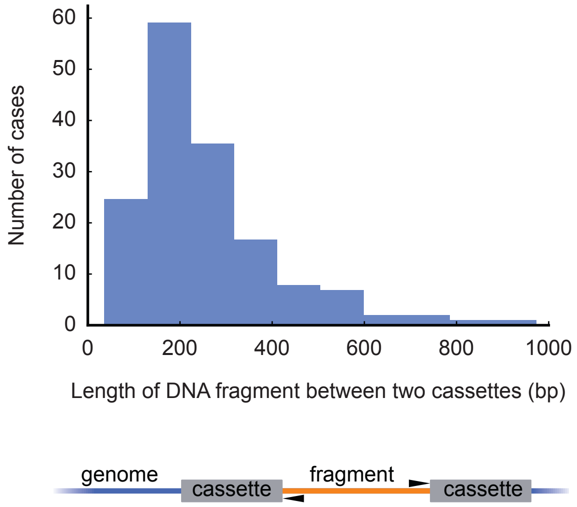
We checked how many of the colonies physically missing from each plate had a flanking sequence mapped to them, and performed the same comparison for all pairs of plates to check for possible accidental plate switches or duplications. The color of each square indicates the percent of physically missing colonies from plate Y that had insertions mapped to them on plate X. On the diagonal, we are comparing each plate to itself, so we expect the number to be low unless a plate was accidentally switched or the deconvolution results are incorrect. Off the diagonal, we expect random values based on how similar the missing colony patterns for the two plates were. A clean diagonal is visible, indicating that our deconvolution method worked and that no plates were switched. Note that numbers on the diagonal >0% can be either deconvolution errors or cases where a colony was present during superpooling but was lost during later library propagation.

Numbers of flanking sequence (►) pairs in different categories:
(in parentheses: percentage of flanking sequence pairs that include a 5' and a 3' flanking sequence, percentage of pairs with both flanking sequences confirmed by LEAP-Seq)

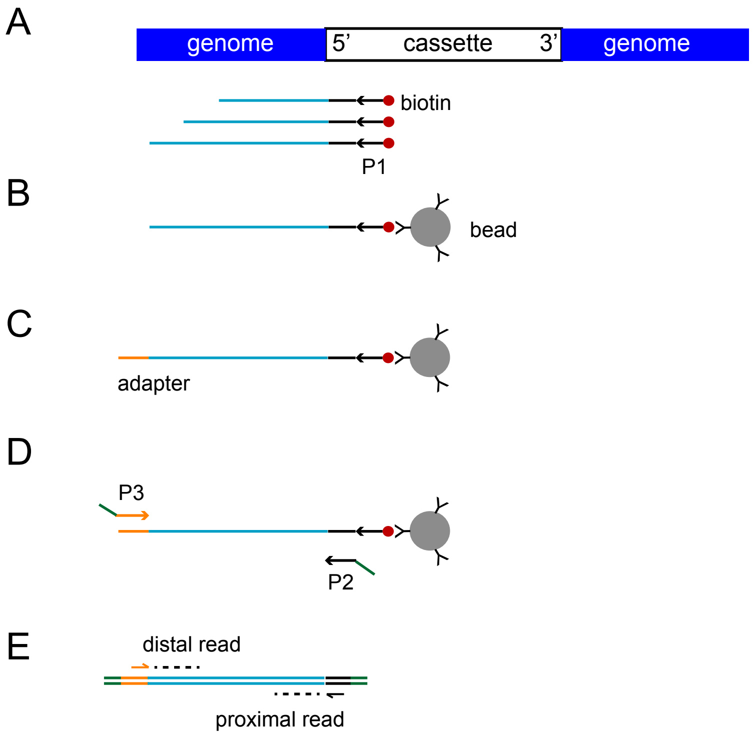


Same chromosome:			**distance between flanking sequences**				
			0 bp	**1-10 bp**	**11-100 bp**	**101-1,000 bp**	**1+ kb**
relative orientation	**same-facing**		-	2 (0%, 0%)	2 (50%, 0%)	0 (0%, 0%)	54 (63%, 3.7%)
	toward-facing		52 (88%, 67%)	52 (87%, 63%)	24 (88%, 79%)	2 (50%, 0%)	29 (55%, 0%)
	away-facing		-	33 (88%, 70%)	14 (57%, 21%)	148 (58%, 0.7%)	27 (70%, 37%)
Different chromosomes:			1166 (68%, 2%)				

Supplemental Figure 5. Pairs of flanking sequences mapped to the same colony can be divided into insertion scenarios based on their distance and relative orientation, cassette sides, and LEAP-Seq confidence data.

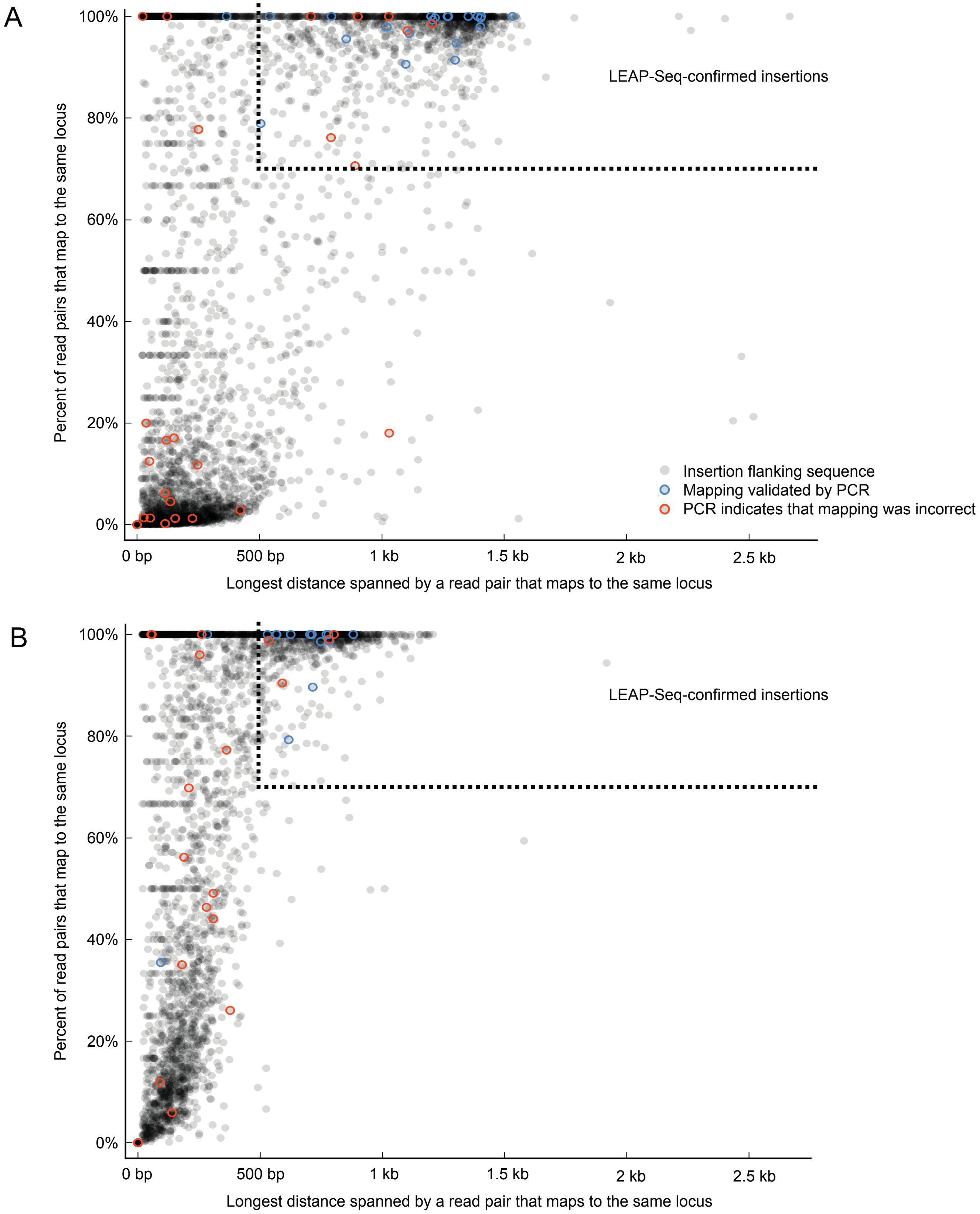


Supplemental Figure 6. The distribution of lengths of putative genomic DNA fragments surrounded by two cassettes was estimated from data shown in Supplemental Figure 5.



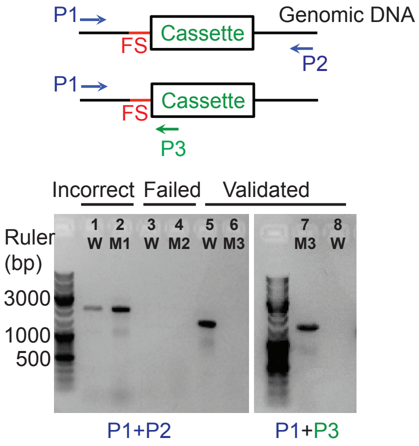
Supplemental Figure 7. LEAP-Seq (Linear and Exponential Amplification of insertion site sequence coupled with Paired-end Sequencing) was employed to obtain longer flanking sequences to confirm cassette insertion sites.

LEAP-Seq was performed on both sides of the insertion cassette. The scheme on the 5’ side is shown here.
(A) A biotinylated primer (P1) binds to the cassette and is extended. Multiple denaturation, annealing and extension cycles are performed. During each cycle, one single-stranded DNA is generated from each cassette.
(B) Products are captured with streptavidin-coated magnetic beads.
(C) A single-stranded DNA adapter is ligated to the end of the captured products.
(D) Ligation products are amplified with a primer binding to the adapter (P3) and another cassette-side primer (P2, closer to the end of the cassette than P1). P2 and P3 have sequences compatible with Illumina sequencing.
(E) Paired-end sequencing is performed to obtain proximal and distal reads (shown as dashed lines) of the flanking sequence.



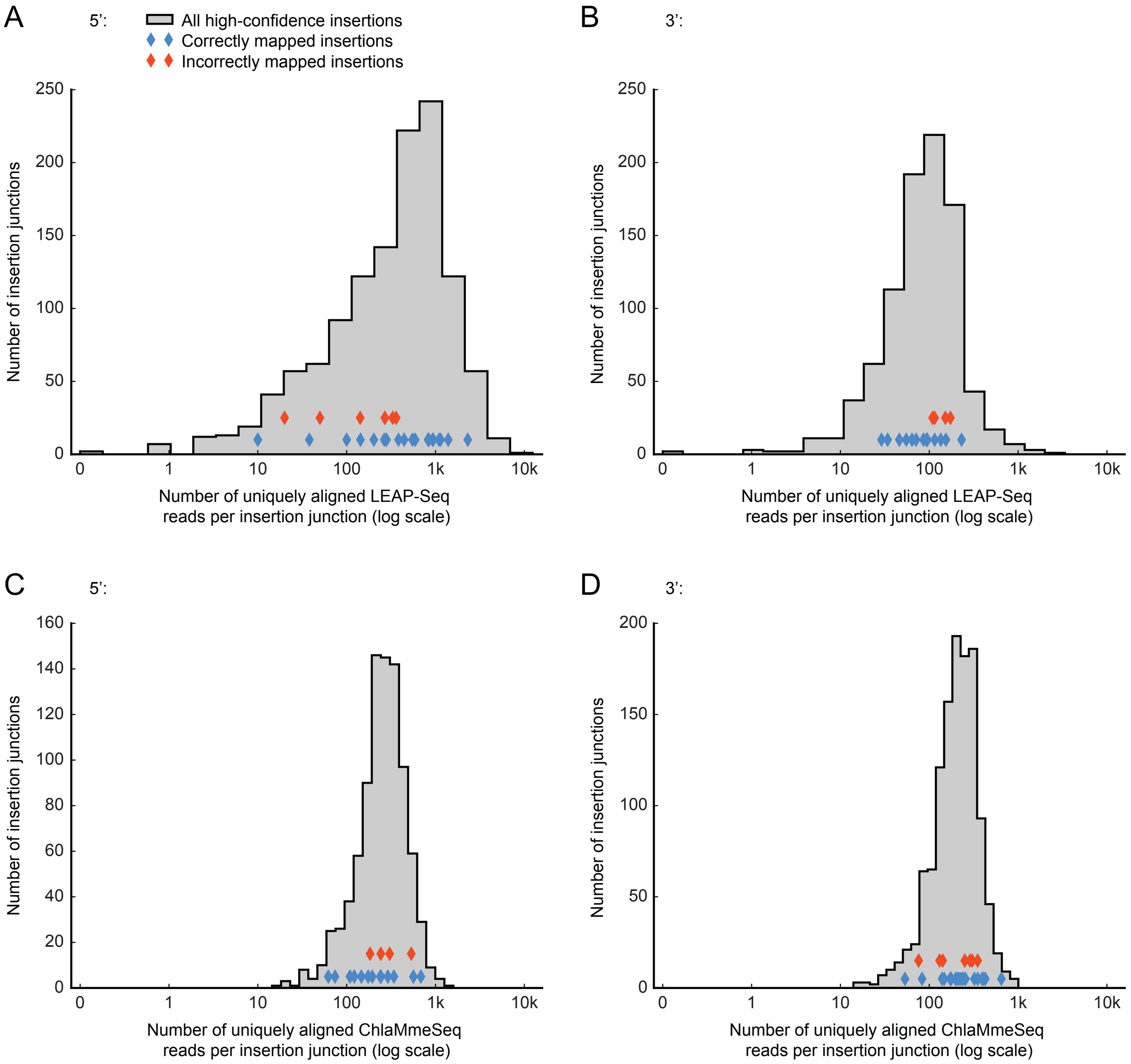
Supplemental Figure 8. Full LEAP-Seq confidence plots for 5' (A) and 3' (B) sides of the cassette, with the same confirmed mutant set cutoffs as in Figure 4B. The blue and red circles show correctly and incorrectly mapped insertions based on PCR validation; the PCR validation results are summarized in Figure 4C and 4D and presented in detail in Supplemental Data Set 9.

The horizontal clusters at Y axis values 50%, 33%, 66%, 25% etc represent insertions with <10 total reads, which were excluded from Figure 4B. The differences between the 5’ and 3’ sides are likely due to higher read lengths achieved in the 5’ data.



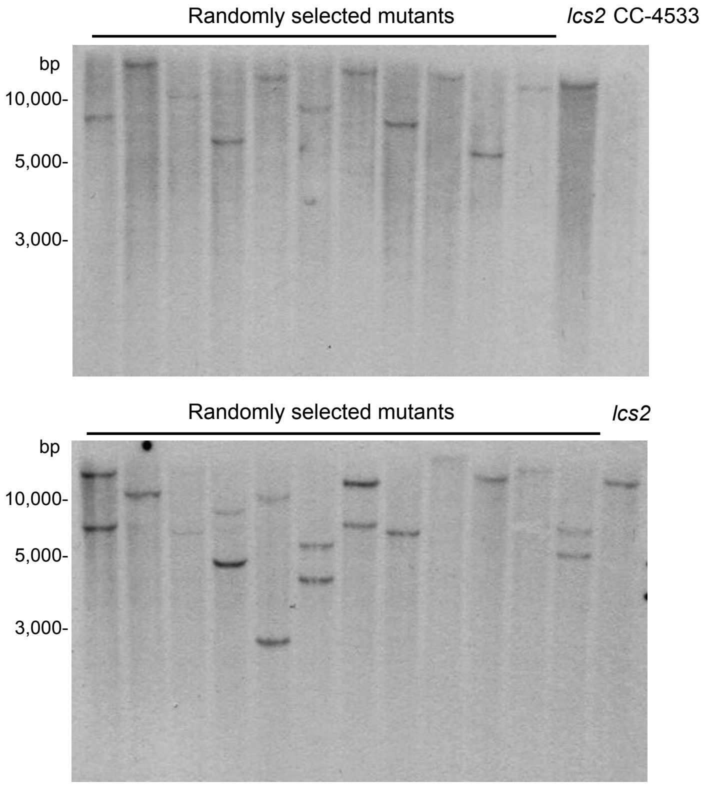
Supplemental Figure 9. The insertion sites of randomly chosen individual mutants were checked by PCR.

Raw data are shown for example mutants. Genomic primers (P1, P2) were designed to anneal approximately 1 kb away from the flanking sequences (FS) extracted by ChlaMmeSeq. Primers P1+P2 were used to amplify the genomic region in both the background strain CC-4533 (WT) and the mutants (M1, or M2, or M3) containing the flanking sequences. If both WT and the mutant produced identical PCR products (lanes 1 & 2, confirmed by sequencing), the insertion site reported by the flanking sequence was considered “incorrect.” If WT produced the expected PCR product but the mutant did not (lanes 5 & 6), the insertion site was further confirmed by PCR across the genome-cassette junction using one of the genomic primers (P1, on the same side of the flanking sequence) and a cassette specific primer (P3). If the mutant produced the expected junction PCR product and the WT did not (lane 7 & 8), the insertion site reported by the flanking sequence was considered “validated by PCR.” For a small number of the randomly chosen mutants, the genomic primers surrounding the flanking sequences did not yield any PCR products in either WT or the mutant (lane 3 & 4) even after two trials, possibly due to differences between our strain's genome and the reference or local sequences that were difficult to amplify by PCR. These cases were grouped as “failed PCR” without further analysis.



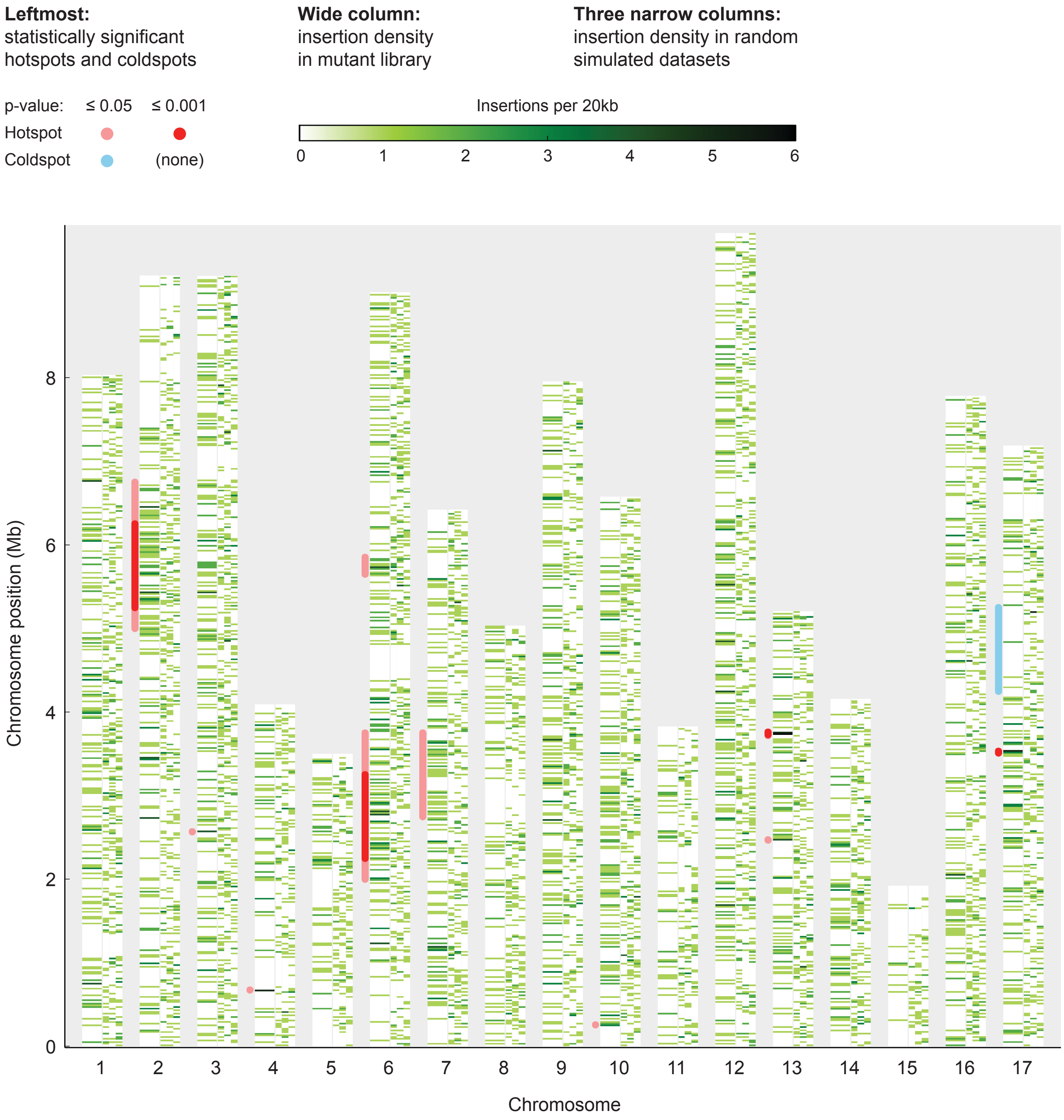
Supplemental Figure 10. The read count in LEAP-Seq and ChlaMmeSeq has little impact on the validation of insertion sites by PCR.

(A-B) The distribution of LEAP-Seq read counts per insertion junction compared to the read counts of insertion sites validated by PCR, for the 5' and 3' sides of the cassette. The read counts for the correct and incorrect validated insertion sites were compared using the Kruskal-Wallis test: the p-values are 0.086 for 5' and 0.079 for 3'.
(C-D) The distribution of ChlaMmeSeq read counts per insertion junction compared to the read counts of insertion sites validated by PCR, for the 5' and 3' sides of the cassette. The p-values, calculated as above, are 0.62 for 5' and 0.31 for 3'.



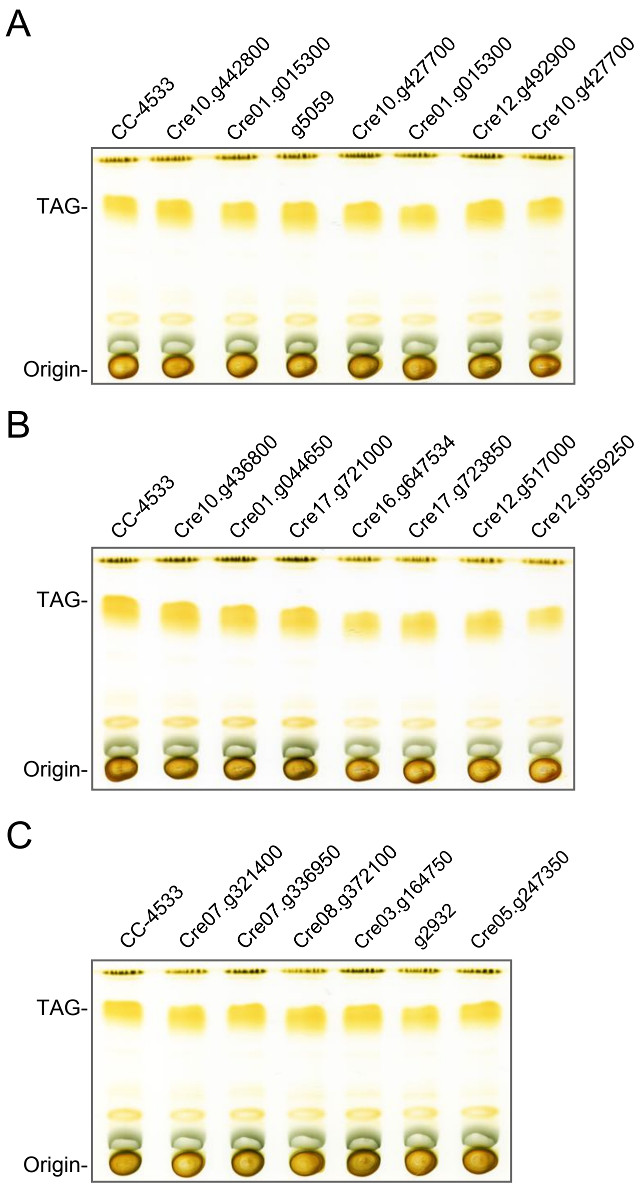
Supplemental Figure 11. 17 out of 23 (74%) of mutants in the LEAP-Seq-confirmed set harbor a single insertion.

lcs2 (discussed later) together with 23 randomly selected LEAP-Seq confirmed mutants were analyzed by DNA gel blotting using the coding sequence of AphVIII as the probe. CC-4533 was included as a negative control.



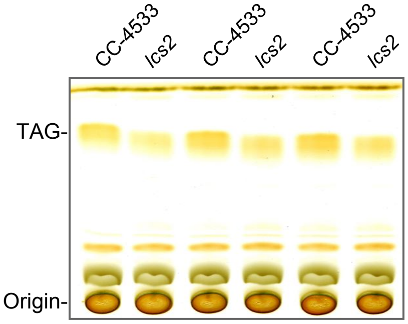
Supplemental Figure 12. The distribution of insertions in the genome is largely random.

For each chromosome, observed insertion density is shown as a heatmap in a wide column, followed by three narrow columns depicting three simulated datasets in which insertions were placed in randomly chosen mappable genomic locations. The simulated data provide a visual guide to the amount of variation expected from a random distribution. The large white areas present in both the observed and simulated data correspond to repetitive genomic regions in which insertions cannot be mapped uniquely. The red and blue lines show potential insertion hotspots and coldspots, which cover a relatively small fraction of the genome. These data are consistent with our earlier observations (Zhang et al. 2014).



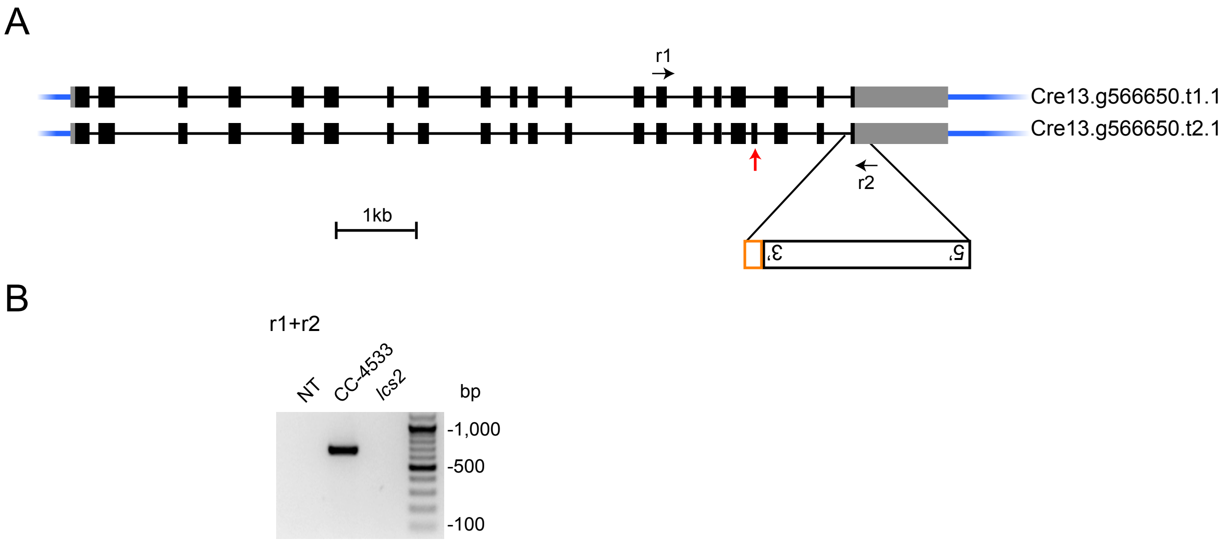
Supplemental Figure 13. Mutants deficient in lipid droplet proteins were screened for TAG deficiency using thin-layer chromatography.

(A-C) The plate containing the *lcs2* mutant is presented in Figure 6B, and the remaining plates are presented here. Lipid extracts from cells with the same amount of chlorophyll were loaded for each mutant (indicated with the gene ID in the v5.3 Chlamydomonas genome). Note that in panel (B), the four mutants on the right side seemed to have visibly lower amount of materials loaded compared to the four samples on the left. Thus, we did not interpret these mutants as candidates with lower amount of TAG accumulated and did not select them for further studies.



Supplemental Figure 14. We confirmed the TAG deficiency in *lcs2* by TLC.

Lipid loading was normalized by chlorophyll content. Three independently grown replicates of CC-4533 and *lcs2* were analyzed.



Supplemental Figure 15. RT-PCR supports the Cre13.g566650.t1.1 model of *LCS2*.

(A) Two gene models exist for *LCS2* in the Phytozome v5.3 Chlamydomonas genome, with one extra exon present in the Cre13.g566650.t2.1 model (indicated by a red arrow).
(B) RT-PCR with primer pair r1 and r2 yielded a single band from CC-4533, which, after sequencing, revealed the absence of the extra exon in the Cre13.g566650.t2.1 model. NT indicates a no-template control.

Supplemental Table 1. Summary of variants of our CC-4533 strain relative to the reference Chlamydomonas v5.3 genome and several other strains sequenced in Gallaher et al., 2015.

	Total Small Variants	High Impact Small Variants	Total Large Variants	High Impact Large Variants
Private	466	3	26	0
Relative to reference	166574	500	1216	242
Relative to CC-503	165033	461	1070	192
Relative to CC-124	71266	213	690	104
Relative to CC-125	164587	452	1031	176
Relative to CC-1690	117059	355	Not Available	Not Available

Small variants were predicted by GATK's UnifiedGenotyper and include single nucleotide substitutions and small insertions and deletions. Large variants were predicted by Pindel and Breakdancer and include large insertions and deletions (> 20 bp) and chromosomal translocations. Overlapping variants from the two sets were mostly removed or tagged (see Supplemental Methods). High impact variants are variants predicted to cause a frameshift or premature translation stop. Private variants are variants found in CC-4533 but not in any of a collection of 39 diverse commonly used laboratory strains (Gallaher et al., 2015). CC-503 is the reference genome strain that was re-sequenced using the same short-read pipeline as the other strains. CC-124 (137c, mt-), CC-125 (137c, mt+) and CC-1690 (21gr, mt+) are commonly used lab strains.

Supplemental Table 2. Details of high-impact variants unique to CC-4533.

Gene ID	Description	Position	Nucleotide Change	Impact
Cre01.g010050		chr01:1843500	GT->G	FRAME_SHIFT
Cre01.g021200	F-box family protein	chr01:3352080	TTGATGGACGAGGCTTTTGGAGA->T	FRAME_SHIFT
g11491		chr11:299416	C->CA	FRAME_SHIFT

Supplemental Table 3. Cryopreserved mutants can be recovered at a success rate greater than 98%.

Recovery of cryopreserved strains					
Experiment 1			Experiment 2		
No. of strains frozen	No. of strains recovered	Recovery rate	No. of strains frozen	No. of strains recovered	Recovery rate
78	77	99%	95	94	99%
83	81	98%	94	94	100%
84	84	100%	92	92	100%
88	88	100%	75	75	100%

For each experiment, four independent 96-well plates containing different strains were cryopreserved and recovered. The numbers of strains frozen and recovered were counted for each plate.

Supplemental Table 4. Sequences of oligonucleotide primers used for LEAP-Seq. All are shown 5’ to 3’.

Primer	Sequence
P1 (5’)	*o*GGCCGTGAGAGGGAGAGC
P1 (3’)	*o*CAGGCCATGTGAGAGTTTGC
Adapter	*P*AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTACTCAGTAGTTGTGCGATGGATTGATG*x*
P2 (5’-MiSeq)	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCTCGTGTGACGATTGGTTCC
P2 (5’-HiSeq)	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGTGTGACGATTGGTTCCAAC
P2 (3’)	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGACGTTACAGCACACCCTTG
P3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Sequencing primer (read 1)	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Sequencing primer (read 2, 5’-MiSeq)	GCTCTTCCGATCTCTCGTGTGACGATTGGTTCC
Sequencing primer (read 2, 5’-HiSeq)	CTCTTCCGATCTCTCGTGTGACGATTGGTTCCAAC
Sequencing primer (read 2, 3’)	CAACACGCCCCGCGCTCC

Modifications: *o*, biotinylation; *P*, phosphorylation; *x*, dideoxycytidine.

Supplemental Table 5. Sequences of oligonucleotide primers used for genotyping and RT-PCR of CC-4533 and *lcs2*. All are shown 5’ to 3’.

Primer	Sequence
g1	CCTCACCTCGCTTCAACTTG
g2	GGTGAAAGGATGTACGTGCC
c1	GCTGGCACGAGTACGGGTTG
c2	GCTCGTGGAGCTCTGAATCT
r1	GCCAACTTCACCGGCTACTA
r2	TGCTACATCCTCTGCTACGG

SUPPLEMENTAL METHODS

Whole-Genome Sequencing Analysis of the CC-4533 Background Strain

Genomic DNA was obtained from CC-4533 by phenol:chloroform:isoamyl alcohol extraction as described previously (Zhang et al., 2014), and was sheared using a Focused-ultrasonicator system (model S2, Covaris, Woburn, MA). The following settings were used to obtain DNA fragments approximately 300 bp in size: intensity: 4, duty cycle: 10, cycles per burst: 200, treatment time: 80 s, temperature: 7°C, water level-S2: 12, water level-E210: 6, sample volume: 130 μL. The fragments were applied onto the SPRIworks System I (Beckman Coulter, Indianapolis, IN) for library preparation and sequenced on a HiSeq platform (Illumina, San Diego, CA). The raw sequences were aligned to the Chlamydomonas version 5 reference sequence with BWA mem (Version 0.7.5a-r405) (Li and Durbin, 2009) using default parameters. Duplicate read pairs were removed using Picard MarkDuplicates Version: 1.85(1345) (http://broadinstitute.github.io/picard), default parameters.

For small variant detection, the Genome Analysis Toolkit Version 2.6-5-gba531bd (McKenna et al., 2010) was used to prepare and call variants on the aligned, de-duplicated reads. Indel realignment, base recalibration, BAQ correction (Li, 2011), removal of heterozygous region variants, and variant quality score recalibration were all performed as previously described (Gallaher et al., 2015). Alignments for 100 randomly selected variants were inspected visually using IGV (Robinson et al., 2011). The functional impact of each variant was predicted using SnpEff (Cingolani et al., 2012).

Mutant Generation and Propagation

The CC-4533 strain was grown in Tris-Acetate-Phosphate (TAP) medium using a modified trace element solution (Kropat et al., 2011) (pH 7.5). Unless otherwise indicated, solid medium plates consisted of 55 mL of TAP medium with 20 μg/mL paromomycin (Sigma-Aldrich, St. Louis, MO), solidified with 1.5% (w/v) agar (MP Biomedicals, LLC, Solon, OH), in rectangular PlusPlates (Singer Instruments, Somerset, UK).

Transformation was performed using *Mly*I-digested plasmid pMJ013b as described previously (Zhang et al., 2014), with one difference: the 16°C incubation was performed for at least 80 min.

After two weeks of growth, the colonies were picked using a Norgren CP7200 Colony Picking Robot (Norgren Systems, Ronceverte, WV). From 375 to 900 colonies were picked from each transformation plate and arrayed into 384-colony format onto solid medium plates.

Libraries were propagated by robotic replication of colony arrays on solid medium plates. Two library copies were kept at room temperature (20-25 °C) in low light (~5 μmol photons m^{-2} s^{-1}), as stacks of 10 plates, each stack in a one-gallon ZipLoc bag (S.C. Johnson & Son, Inc, Racine, WI), and all bagged stacks in a plastic bin. Each copy was propagated every four weeks and the replication of one copy was staggered relative to the other by two weeks. This protocol provides redundancy that protects against colony loss resulting from incorrect media preparation or large-scale contamination. A RoToR robot (Singer Instruments, Somerset, UK) was used for replica plating with the following parameters: Source and Target Pinning pressure: 30%, speed: 19 mm/s, overshoot: 2 mm, repeat pin: 1x, Dry Mix clearance: 0.5 mm, diameter: 1 mm, 3 rotation cycles.

Colonies lost and gained were identified by image analysis software and confirmed visually. The analysis compared images of one copy of the library plates obtained on 12/17/2012 with images obtained 18 months later on 6/20/2014. Colony presence and absence at each time point was evaluated with the HT Colony Grid Analyzer Java program (Collins et al., 2006). Lost or gained colonies identified in this manner were visually confirmed.

Cryopreservation

Our cryopreservation protocol is based on a procedure previously developed for individual strains (Crutchfield et al., 1999). Cells were inoculated from each 384-colony agar plate into four 96-well conical bottom plates (651261, Greiner Bio-One, Monroe, NC) containing 120 μL of TAP per well by using the RoToR "4:1

Breakdown" program. Settings were as follows: Source Pinning pressure: 10%, speed: 19 mm/s, overshoot: 2 mm, repeat pin: 1x, Dry Mix clearance: 0.5 mm, diameter: 1 mm, 3 rotation cycles, Target Pinning speed: 19 mm/s, backoff: 0.5 mm, repeat pin: 1x, Wet Mix diameter 1 mm, speed 25 mm/s, 5 rotation cycles, travel 3D: 5. 96-well plates were grown in low light (~2 μmol photons m^{-2} s^{-1}) to a density of ~2.0 x 10^6 cells/mL. 100 μL of TAP with 5% (v/v) methanol was then added to each well. Samples were mixed and sealed with adhesive foil covers (89049-034, VWR). Plates were slowly cooled in styrofoam boxes in a -80°C freezer for 90 minutes and then moved into the vapor phase of a MVE 815P-190 liquid nitrogen freezer (Chart Industries, Ball Ground, GA) for long term storage.

For recovery, 96-well conical bottom plates were removed from the liquid nitrogen tank and immediately thawed in a 30°C water bath for 8 minutes. Plates were centrifuged in an Eppendorf 5810 centrifuge (Eppendorf, Hamburg, Germany) for 5 minutes at 500 x *g*. The supernatant was removed by aspiration and the pellet was resuspended in 100 μL of fresh TAP medium. The plates were stored for 3 h low light (~2 μmol photons m^{-2} s^{-1}) at room temperature (20-25°C). The cells were pelleted again by centrifugation for 5 min at 500 x *g* and the supernatant was replaced with 100 μL of fresh TAP medium. Cells were then grown in low light until they reached a density of 1.5 x 10^6 cells/mL (approximately two weeks). The plates were centrifuged for 5 minutes at 500 x *g*, the supernatant was removed and the pellets were pinned with the RoToR robot onto agar plates in a 96-colony array format.

Combinatorial Super-pooling Scheme

Our super-pooling designs are based on binary error-correcting codes. An [N,K,D] binary error-correcting code is one that encodes all possible K-length binary strings into N-length binary strings with a minimum Hamming distance of D between each pair of encoded strings, resulting in a set of 2^K N-length binary strings that can be used as combinatorial super-pool signatures.

The existence of different error-correcting codes was checked using an online database (http://www.math.unl.edu/~djaffe2/codes/webcodes/binary/codeform.html). Based on those data, we chose codes of sufficient size for our purposes, that had at least 6 bits of difference between each codeword: the [15,6,6] code for 53 plate-pools, and the [20,10,6] code for 384 colony-pools. The generating matrix for the [15,6,6] code was obtained from the database above; the [20,10,6] code was generated by adding a parity bit to the [19,10,5] code (generating matrix obtained from (Betsumiya et al., 2003)) using the add_parity_bit function in binary_code_utilities.py.

For the purpose of sister colony detection for colony-pools, we took the subset with a bit sum of 8-10 (588 codewords) from the [20,10,6] code, using the choose_codewords_by_bit_sum function. It was checked for sister colony conflict possibilities using the clonality_conflict_check function: no conflicts were detected up to 2 errors, meaning there are at least 3 bits of difference between every codeword and the sister colony result of any other two codewords.

The final code subsets corresponding to the numbers of plate-pools and colony-pools were chosen to minimize the difference between the number of samples in each super-pool, using the give_N_codewords_even_distribution function with 100 random trials and picking the best result: for plate super-pooling, 53 codewords were chosen from the [15,6,6] code, and for colony super-pooling, 384 codewords were chosen from the 8-10 bit-sum subset of the [20,10,6] code. The final codeword lists are illustrated in Figures 2B and 2C, and provided as Supplemental Data Sets 3 and 4; all the functions listed by name are part of the binary_code_utilities.py program, which was uploaded to GitHub: (http://github.com/Jonikas-Lab/Li-Zhang-Patena-2015).

Pooling and Flanking Sequence Extraction

To generate pools by plate, libraries were grown on agar in low light (~5 μmol photons m^{-2} s^{-1}) for 1 week. Colonies from each plate were scraped and resuspended in 5 mL of fresh TAP medium, then vortexed for 3 min to produce each plate-pool. 2 mL of each plate-pool was transferred to a specific well of a 96 deep-well plate (USA Scientific, Ocala, FL).

To generate pools by colony, libraries were grown on agar in ~5 μmol photons m^{-2} s^{-1} for 1 week. Pooling was accomplished with the RoToR robot using the 4:1 Breakdown program with the same RoToR parameters as above. Cells from each 384-colony agar plate were transferred into the same set of four 96-well plates containing 200 μL TAP. The pooled samples were then transferred from the four 96-well shallow plates to four 96-well deep plates for super-pooling.

Super-pooling of both by-plate and by-colony pools was accomplished with a BioMek liquid-handling robot (Beckman Coulter, Indianapolis, IN). Pools were combined according to the schemes in Supplemental Data Sets 3 and 4, by transferring 10 μL of each pool sample to the appropriate super-pool sample. Twenty super-pools were made for pools by colony, and fifteen for pools by plate. ChlaMmeSeq was performed on each super-pool as described previously (Zhang et al., 2014). The products were sequenced on the Illumina Genome Analyzer IIx platform at the Stanford Sequencing Service Center.

Flanking Sequence Deconvolution

The ChlaMmeSeq data from each super-pool were analyzed as in (Zhang et al., 2014): genomic flanking sequences were grouped together based on genomic alignment position and orientation; no low-abundance flanking sequences were removed, and no adjacent positions were merged. Flanking sequences derived from the 5' and 3' cassette sides were analyzed separately. The read counts for all flanking sequences in all super-pools were normalized to reads per million. The normalized read counts are provided in Supplemental Data Set 5.

Then the read counts were converted to 0/1 values denoting the absence/presence of each flanking sequence in each super-pool with a method designed to tolerate minor cross-contamination between super-pools and other types of noise, illustrated in Supplemental Figure 2. The method uses 3 parameters (N, m, x), which were optimized separately for different data sets (plate and colony, 5’ and 3’). It was applied separately to each flanking sequence to allow for mapping of low-abundance as well as high-abundance flanking sequences. The method is as follows: (1) The Nth highest normalized read count for that flanking sequence is found – that value is called R. (2) If R is below m, the flanking sequence is discarded as low-abundance noise. (3) The flanking sequence is marked as present in super-pools in which it had at least x*R reads, and absent in the remaining super-pools – this creates a binary observed codeword for that flanking sequence. (4) The observed codeword is compared to the codewords used for each plate-pool or colony-pool during combinatorial pooling; if there is a single closest codeword with at most 2 errors (i.e. differences between the observed codeword and the expected codeword for the pool), the flanking sequence is considered mapped to the corresponding pool, and otherwise it is considered unmapped.

This method was applied separately to the colony-super-pool and plate-super-pool data, from the 5' and 3' sides of the cassette, using 90 different combinations of parameters in order to optimize the number and quality of mapped flanking sequences (N values of 4,5,6 for plate data and 6,7,8 for colony data, since each pool was included in at least 6 plate-super-pools and at least 8 colony-super-pools; m values of 10, 20, 30, 50 and 80 reads per million; x values of 0.03, 0.05, 0.07, 0.1, 0.2 and 0.3). We decided to apply three iterations of deconvolution to each data set: iteration 1 used parameters optimized for the highest fraction of 0-error results out of all mapped flanking sequences to give the highest quality and the lowest number of false positives; iteration 2 used parameters optimized for the highest fraction of results with 0 or 1 error; iteration 3 used parameters optimized for the highest total number of results with 0-2 errors (the combinatorial pooling scheme we used guarantees 6 bits of distance between all pairs of codewords, so 2 errors are the highest number that can reliably be corrected because observed codewords cannot be less than 3 errors away from two different pool codewords). The optimized parameters used for all three iterations for each data set are given in Supplemental Figure 2B. The three iterations were applied in sequence, first to take the highest-quality set of mapped flanking sequences, and then to add more sequences of increasingly lower mapping quality: flanking sequences that had 0 errors during iteration 1 were marked as mapped; out of the remaining flanking sequences, ones that had

0-1 errors during iteration 2 were marked as mapped; out of the remaining flanking sequences, ones that had any unique best match in iteration 3 were marked as mapped. For each sequence, we kept track of both which mapping iteration it was included in and how many mapping errors it had. The resulting numbers of mapped flanking sequences are shown in Supplemental Figure 3.

Verification of Deconvolution Results

To verify the quality of deconvolution and detect potential major errors in combinatorial pooling, we took advantage of the fact that most plates have 10-50 blank positions distributed in random locations. If the mapping process was done correctly, we would expect that few sequences would be assigned to blank positions. Indeed, only 1.3% of blank positions in the library were assigned flanking sequences (probably due to these positions originally containing colonies which were lost during propagation), in comparison to 38% of positions containing colonies (Supplemental Figure 4). We conclude that no major errors occurred in the position mapping protocol.

To quantitatively evaluate the reliability of deconvolution, two non-overlapping sets of 96 mutants were randomly chosen among the LEAP-Seq confirmed mutants (described below). The mutants in each set were grown under ~10 μmol photons m^{-2} s^{-1} light in liquid TAP medium, initially in 96-well plates (one mutant per well) and then combined into a single pooled culture in an Erlenmeyer flask. Cells were harvested at 3-8 X 10^6 cells/mL. ChlaMmeSeq and data analysis were performed on each set of mutants as described previously (Zhang et al., 2014) except that the resulting PCR products were sequenced on an Illumina MiSeq platform. The flanking sequences present in the data were compared to the flanking sequences expected to be present in the picked mutants based on the deconvolution results: 98 out of 103 expected flanking sequences were found in the first set of mutants, and 90 out of 103 in the second set, yielding a verification rate of 95% and 87% respectively. Nearly all of the present and expected sequences had over 1k reads, so the missing sequences are unlikely to be absent due to low sequencing coverage.

LEAP-Seq (Linear and Exponential Amplification of Insertion Site Sequence Coupled with Paired-End Sequencing)

Equal-volume aliquots from each plate-pool were combined to obtain the whole library pool. Genomic DNA was extracted from the pool and used as the template for primer extension. Each 50 µL primer extension reaction contained 500 ng of genomic DNA, 10 μL Phusion GC buffer, 2 μL of 0.25 µM biotinylated primer P1 (different for 5' and 3'; sequences of oligonucleotides used for LEAP-Seq are in Supplemental Table 4), 3 μL dimethyl sulfoxide (DMSO), 1 μL 50 mM MgCl$_2$, 1 μL deoxynucleotide triphosphate mixture (dNTPs; 10 mM for each) and 0.5 μL Phusion Hotstart II (HSII) polymerase (F549L, Thermo Fisher Scientific). The reaction mixtures were incubated at 98°C for 3 min followed by 40 thermal cycles of 10 sec at 98°C, 30 sec at 65°C and 18 sec at 72°C. The incubation and thermal cycles were repeated for one round after an additional 0.5 μL of polymerase was added. Reaction mixtures were then heated at 95°C for 5 min, chilled on ice water for at least 1 min and kept at 4°C before bead binding.

Bead binding was performed with reagents in the Dynabeads kilobaseBINDER Kit (60101, Life Technologies, Carlsbad, CA). 10 μL streptavidin-coupled beads from the kit were added to a fresh set of tubes. The beads were washed twice with 100 μL PBS (SH3025601, Thermo Fisher Scientific) and then washed once with 20 μL of the binding buffer supplied with the kit. For each wash, the beads were immobilized at the bottom of the tube using the DynaMag magnet (12331D, Life Technologies), the supernatant was discarded and the beads resuspended in fresh buffer. After the final wash, beads were resuspended in 100 μL of binding buffer. The primer extension mixtures kept at 4°C were diluted by addition of 50 μL H$_2$O, and then were added to the resuspended beads. The tubes were then incubated 15-18 h on a rotary shaker at room temperature.

For ligation, the beads were collected on the DynaMag, the supernatant was removed, then the beads were washed 3 times with 100 µL PBS and finally resuspended in a 20 μL ligation cocktail. The ligation cocktail contained 11.25 μL H_2O, 4 μL betaine, 2 μL CircLigaseII reaction buffer, 1 μL 25 μM single-stranded adapter, 1 μL 50 mM $MnCl_2$ and 0.75 μL CircLigaseII (CL9025K, Epicentre, Madison, WI). This reaction mixture was immediately placed on a thermocycler block pre-adjusted to 60°C, and incubated at 60°C for 1 h.

For exponential PCR, the beads were precipitated, washed 3X with PBS as described above and resuspended in 50 μL exponential PCR cocktail. The PCR cocktail included 32.5 μL H_2O, 10 μL Phusion GC buffer, 3 μL DMSO, 1 μL 50 mM $MgCl_2$, 1 μL 10 mM dNTPs, 1 μL 25 μM primer P2 (different for 5’ and 3’ samples), 1 μL 25 μM primer P3 and 0.5 μL Phusion HSII polymerase. Reaction mixtures were incubated at 98°C for 3 min, followed by 10 three-step thermal cycles (10 sec at 98°C, 30 sec at 63°C and 20 sec at 72 °C) and then 10 two-step thermal cycles (10 sec at 98°C and 45 sec at 72°C).

Five to eight tubes of exponential PCR products were combined and concentrated using a MinElute Kit (28006, Qiagen, Venlo, Netherlands). The concentrated products were resolved on agarose gels and DNA fragments in the range of 750-1,500 bp were gel purified using a MinElute Kit. For each side of the cassette (5’ and 3’), LEAP-Seq was performed twice, with the products analyzed by Illumina MiSeq and HiSeq paired-end sequencing.

LEAP-Seq Data Analysis

The LEAP-Seq samples were sequenced using two methods: Illumina HiSeq paired-end 100bp reads, and Illumina MiSeq paired-end 36bp reads. The sequencing was done separately for 5' and 3' cassette flanking sequences, with 14M HiSeq 5' reads, 25M HiSeq 3' reads, 2.6M MiSeq 5' reads, and 3.7M MiSeq 3' reads. In both cases, the raw data consisted of a proximal read (containing some cassette sequence and the genomic flanking sequence) and a distal read (containing a more distant genomic sequence). Data were processed as follows:

(1) The initial cassette sequence was removed from the beginnings of the proximal reads, allowing 20% errors; reads that did not contain the expected sequence were discarded. For MiSeq data, the initial cassette sequence was

only 3 bp, too short to allow any errors, so it was simply trimmed using the command "deepseq_preprocessing_wrapper.py -F AAC -A NONE"; 22% of the 5' reads and 49% of the 3' reads contained the expected sequence. For HiSeq data, the initial cassette sequence was 21 bp in length for the 5' end and 71 bp in length for the 3' end. Both HiSeq samples were trimmed using cutadapt version 1.2rc2 (Martin, 2011) using the command "cutadapt -g ^<seq> -e 0.2 --untrimmed-output=/dev/null", with <seq> being the expected cassette sequence (CGTGTGACGATTGGTTCCAAC for the 5’ data and GACGTTACAGCACACCCTTGATCATCATCAGCTGCTCTTCCCTGCCGCTGCAACACGCCCCGCGCTCCAAC for 3’). 86% of 5' reads and 17% of 3' reads contained the expected sequence.

(2) The proximal sequences were trimmed to the initial 21 bp using the command "fastx_trimmer -l 21 -Q33" (http://hannonlab.cshl.edu/fastx_toolkit), since ChlaMmeSeq data only yields 21-bp flanking sequences. Distal sequences were similarly trimmed to 30 bp, in order to make the alignment easier.

(3) The reads were aligned to the Chlamydomonas genome and to the cassette sequence, using bowtie version 1.0.0 (Langmead et al., 2009) with “-v1 -k10 --strata --best --tryhard” options (allowing up to one mismatch; bowtie doesn't allow indels). Two separate alignment runs were done for each sample: one against our insertion cassette (GenBank accession number KJ572788, insertion cassette feature only), and one against the Chlamydomonas v5.3 genome (from Phytozome, no repeat masking) with the chloroplast and mitochondrial genomes added as separate chromosomes (from the National Center for Biotechnology Information (NCBI) website: NC_005353 and NC_001638). Both result files were parsed in parallel to categorize the reads: reads that did not align to either reference were counted and discarded; reads that aligned to the cassette were categorized as cassette-aligned, regardless of any genome alignment; reads that aligned uniquely to the genome categorized as genome-aligned; reads that aligned to multiple genomic locations were counted and discarded. A custom wrapper script deepseq_alignment_wrapper.py was used to accomplish these steps.

(4) The genomic and cassette alignment files for the proximal reads were parsed (in python, using HTSeq, http://www-huber.embl.de/users/anders/HTSeq), the bowtie flanking sequence alignment locations were converted to the cassette insertion locations (by adjusting the position and strand based on which end of the flanking region was adjacent to the cassette), and all reads with the same insertion position and orientation were joined into single flanking sequences. The command used was "mutant_count_alignments.py --Carette -e 5prime -r reverse" for 5' data and "mutant_count_alignments.py --Carette -e 3prime -r forward" for 3' data; the HiSeq and MiSeq data were merged at this point. Information about read IDs was kept to enable adding the distal reads to the data.

(5) Distal reads and alignment positions were added to the data structures of their matching proximal reads using the mutant_Carette.Insertional_mutant_pool_dataset_Carette.add_Carette_genome_side_alignments_to_data function in python.

(6) Annotation files from the Phytozome bulk downloads page for the Chlamydomonas v5.3 genome were used to get the gene and feature positions (Creinhardtii_236_gene.gff3, parsed in python using BCBio.GFF, http://github.com/chapmanb/bcbb/tree/master/gff) and gene annotation information (Creinhardtii_236_annotation_info.txt). The distal read alignment locations were matched to genes to get the gene ID, feature (exon/intron/UTR), and orientation (sense or antisense compared to gene direction) for each sequence. This was done in python using mutant_Carette.Insertional_mutant_pool_dataset_Carette.find_genes_for_mutants and mutant_Carette.Insertional_mutant_pool_dataset_Carette.add_gene_annotation functions.

(7) The resulting data set was filtered to exclude any proximal flanking sequence alignment positions that were not present in the plate and colony deconvolution results.

(8) For each flanking sequence, the number of read pairs with the proximal and distal read mapping to the same locus, and the highest distance between

such read pairs were calculated with mutant_Carette.Insertional_mutant_Carette methods Carette_max_confirmed_distance, Carette_N_confirming_reads and Carette_N_non_confirming_reads functions.

Complete data for the flanking sequences that were mapped during deconvolution are provided in Supplemental Data Sets 7 and 8.

Analysis of Flanking Sequence Pairs Mapped to the Same Colony

Flanking sequence pairs mapped to the same colony were analyzed both to determine which pairs were derived from two ends of the same insertion versus two independent insertions, and to use the pairs derived from two ends of one insertion to determine the characteristics of the insertion sites (such as presence and size of genomic deletions).

To standardize the analysis, the sets of flanking sequences from each colony with 2+ flanking sequences were subdivided into pairs, generating N-1 pairs from each set of N flanking sequences (2 flanking sequences = 1 pair, 3 = 2 pairs, 4 = 3 pairs, etc.): all the flanking sequences mapped to each colony were sorted by genomic position, and each adjacent pair was taken for pair analysis. This method preserves the total counts of various possible insertion events: for instance, one colony with three flanking sequences all derived from separate insertions would yield two pairs categorized as separate insertions, i.e. two "additional" insertions in one mutant. The pairs were categorized based on their relative orientation and distance, without regard to which side of the cassette they were derived from, because a single insertion can consist of two cassettes in opposite orientations and thus yield two 5' or two 3' ends (we base this on the observation of frequent cases of either end of the cassette appearing as an insertion flanking region, which implies an insertion of two adjacent cassettes – see (Zhang et al., 2014). The analysis is based on three possible relative orientations between two flanking sequences: (1) "same-facing," where the cassette-adjacent sides of both flanking sequences face in the same direction, (2) "toward-facing," where the cassette-adjacent sides of the two flanking sequences face toward each other, and (3) "away-facing," where the cassette-adjacent sides face away from each other. Each category was further subdivided based on the distance between the cassette positions in the two flanking sequences: 0 bp, 1-10 bp, 11-100 bp, 100-1000 bp, and 1+ kb. Note that a distance of 0 bp is impossible for "same-facing" pairs, because the two sequences would be identical and thus would be treated as a single flanking sequence in the basic analysis; also, a distance of 0 bp cannot be properly categorized as "toward-facing" or "away-facing," so the "toward-facing" category was chosen arbitrarily. The relative orientations can only be determined for flanking sequence pairs aligned to the same chromosome; thus, most pairs were categorized as "different chromosomes" with no further sub-categorization based on orientation or distance. The resulting counts in each category are shown in Supplemental Figure 5.

The differences between the numbers of cases of relative orientations within the same distance range were used to determine whether the sequences in a given category are likely to be derived from one or two insertions. For flanking sequence pairs derived from two insertions, the relative orientation would be random and unbiased, yielding 50% same-facing pairs and 25% each of toward-facing and away-facing. Thus, if the numbers in any distance range are significantly different from this expected 2:1:1 distribution, the pairs in the category that has more pairs than expected are likely to be derived from two sides of a single insertion. The findings can be confirmed by two additional factors, shown in parentheses in Supplemental Figure 5: (1) The number of the pairs that are derived from one 5' and one 3' cassette side versus two same sides, because flanking sequences from independent insertions should have an equal chance of being from same or different sides, whereas two sides of a single insertion should be biased toward different sides, although single insertions of two cassette copies in reverse orientations can yield two 5' or two 3' sides; (2) The number of pairs that have both flanking sequences in the LEAP-Seq-confirmed set.

Based on the data in Supplemental Figure 5, we observed several distinct categories of flanking sequence pairs: (1) the 0-distance "toward-facing" pairs are exactly what you would expect from two sides of a single clean insertion; this is corroborated by the high percentage of 5'+3' cases, and the high fraction of LEAP-Seq-confirmed sequences on both sides. (2) The numbers of "toward-facing" pairs in the 1-10 bp and 11-100 bp distance categories are significantly higher than the numbers of "same-facing" pairs in the same distance categories, implying that those are also derived from single insertions. They are likely to be two ends of insertions with a genomic deletion—again, this is confirmed by the high percentage of 5'+3' cases and high-confidence sequences. Cases indicating short deletions may also be clean insertions having PCR or sequencing errors. (The two "toward-facing" pairs in the 101-1000 bp distance bin are likely to be insertions with genomic deletions as well, although the number is too low for differences to be significant. Their distances are 201 bp and 262 bp, and they were included in the "insertions with deletion" category in Figure 3C to avoid artificially excluding potential larger deletions.) (3) The number of "away-facing" pairs in the 1-10 bp category is also much higher than of "same-facing" pairs with the same distances, and again they have a high fraction of 5'+3' and high-confidence cases. They are most likely derived from two ends of a single insertion with a genomic duplication on the two sides of the cassette, perhaps caused by a single-stranded overhang in the genomic break position. Again, cases with short duplications may be clean insertions with PCR or sequencing errors. (4) The number of "away-facing" pairs in the 101-1000 bp category is also much higher than of "same-facing" pairs with the same distances, implying they are derived from a single insertion. However, unlike previous such categories, they have an average percentage of 5'+3' cases and an extremely low percentage of LEAP-Seq-confirmed cases, which implies that the flanking sequences are unlikely to reflect the true position of the insertion.

Given the orientation of those cases, and our previous experience with multi-fragment insertions, our interpretation is that these are most likely cases of a single insertion consisting of a genomic DNA fragment between two cassettes, unrelated to the surrounding genomic DNA on the other sides of the cassettes, with the pair consisting of the two “inner” cassette flanking sequences on two sides of the genomic DNA fragment. The 14 "away-facing" pairs in the 11-100 bp category are likely to be split between cases (3) and (4) – we looked at them in detail and concluded that three (with distances of 11 bp, 12 bp and 19 bp) are likely to be category 3 duplications (since they are LEAP-Seq-confirmed on both sides), while the remaining 11 (with distances 36 bp and higher) are likely to be cassette-fragment-cassette cases in category 4. (5) At the 1+ kb distance, the numbers of "same-facing", "away-facing" and "toward-facing" pairs approximately match the 2:1:1 ratio expected from unrelated positions, and few of them are high-confidence, implying that most of them are derived from two separate insertions or from single insertions with a genomic DNA fragment from a distant locus on one side. The same is the case for the "different chromosome" pairs. Additionally, the two "same-facing" pairs in the 1-10 bp distance category were analyzed in more detail: they both have distances of 1 bp and are in positions with 10 consecutive G bases in the genome, meaning that they are most likely derived from the same original flanking sequences via 1-bp indels during PCR or sequencing—thus, they were not counted in the analysis. The two “same-facing” cases in the 11-100 bp category are likely real cases of two nearby insertions. To produce the estimated numbers of pairs in each category in Figure 3C, half of the number of "same-facing" cases in each distance category was subtracted from the "toward-facing" or "away-facing" cases, because that number is likely to occur by chance from cases of two independent insertions.

To estimate what fraction of all the category 5 pairs are derived from genuine cases of two insertions in a mutant (or the indistinguishable case of two mutant strains in one colony) versus from two sides of a single insertion with a genomic DNA fragment from a distant locus on one or both sides, we looked at the fraction of pairs consisting of 5'+3' ends, versus pairs consisting of 5'+5' or 3'+3' ends. In flanking sequence pairs derived from two independent insertions, the two ends should be independent, thus the fraction of 5'+3' cases should be close to what is expected randomly based on the overall number of 5' and 3' ends (which is 49%). In flanking sequence pairs derived from two ends of one insertion, the fraction of 5'+3' cases should be higher, close to the 88% value observed in the category 1-3 cases which we are confident are two ends of one insertion; the 12% of cases where one insertion has two 5' or two 3' ends are derived from insertions of two adjacent cassettes in opposite orientations, which have been observed in our previous experiments (Zhang et al., 2014). 68% of the category 5 cases are 5'+3', implying that approximately 51% of them are derived from two insertions, and 49% from one insertion, since 51%*49% + 49%*88% = 68%.

Verification of Insertion Sites by PCR

To estimate the fractions of insertion sites that can be validated or identified as incorrectly mapped (Figure 4C and 4D), 92 insertions that mapped to genes were chosen at random from the library (47 confirmed by LEAP-Seq and 45 not confirmed by LEAP-Seq). The mutants carrying these insertions were streaked to single colonies and grown on solid TAP medium with paromomycin in low light (<5 μmol photons m^{-2} s^{-1}). To extract DNA, a single colony of each mutant was picked and resuspended in 50 μL of 10 mM EDTA (pH 8.0) in a well of a 96-well plate. The plate was vortexed for 10 sec, heated to 100°C for 10 min and then cooled to 4°C over a period of 1 min. The plate was then centrifuged at 1,000 x *g* for 1 min to pellet cell debris, and the supernatant from each well containing the genomic DNA was transferred to a fresh microtitre plate and used as template for PCR. PCRs were performed as described previously for characterizing insertion sites (Zhang et al., 2014). Primers were designed to anneal 1-1.3 kb away on each side of the insertion site indicated by alignment of the 20–21 bp ChlaMmeSeq flanking sequence to the genome (version 5.3 from Phytozome, see sequences of primers used for each check PCR in Supplemental Data Set 9).

Verification PCRs were performed in two phases (Supplemental Figure 9): (1) PCR was performed using the genomic primers in an attempt to produce a product across the expected insertion site. If both CC-4533 and the mutant produced the same amplicon spanning the expected insertion site (confirmed by sequencing), the insertion was categorized as "incorrect." If CC-4533 produced the expected PCR band but the mutant did not produce it, we proceeded to the second step: (2) Amplification of the genome-cassette junction from the side that produced the ChlaMmeSeq flanking sequence, with one primer binding to the cassette and the other primer binding to flanking Chlamydomonas genomic DNA ~1kb away. The PCR product, which should contain the cassette-genome junction, was sequenced. If a mutant produced no amplification product in step 1 but yielded an expected PCR product in step 2, the corresponding insertion was categorized as “validated by PCR." In three mutants containing LEAP-Seq-confirmed insertions and seven mutants containing insertions not confirmed by LEAP-Seq, genomic primers surrounding the site of insertion did not yield any PCR products in WT or the mutant even after several trials, possibly due to incorrect reference genome sequence or local PCR amplification difficulties. These cases were grouped as “failed PCR” and were not further analyzed. Given that these three mutants could not be analyzed because of problems with the primers or PCR of that genomic locus in CC-4533, rather than a problem specific to correctly- or incorrectly-mapped mutants, we feel that it is appropriate to exclude these mutants from calculations of the percentage of correctly mapped mutants.

DNA Gel Blotting

5 μg of genomic DNA from phenol:chloroform:isoamyl alcohol extracted (Zhang et al., 2014) nucleic acid was digested with 10X *Stu*I enzyme (New England Biolabs, Ipswich, MA) overnight at 37°C. The resulting fragments were resolved on a 0.7% TAE agarose gel (w/v) at 30 V overnight at 4°C and then for 2 additional h at 100 V. The gel was transferred onto a Zeta-probe membrane (162-0196, Bio-Rad) using the alkaline transfer protocol given in the manual of the membrane. The membrane was rinsed in 2X saline-sodium citrate (0.3 M NaCl, 0.03 M sodium citrate, pH 7.0) and cross-linked twice using the Stratalinker 1800 (Stratagene, La Jolla, CA). Probing and generation of the *AphVIII* probe were performed according to the Amersham Gene Images AlkPhos Direct Labeling and Detection System (RPN3690, GE Healthcare, Little Chalfont, UK) protocols, except for the following changes: the primary washes were at 65°C for 1 h and then 65°C for 2 h, and secondary washes were at room temperature for

2 h and then room temperature for 30 min. PCR primers used for probe generation were oMJ588 (GACGACGCCCTGAGAGCCCT) and oMJ589 (TTAAAAAAATTCGTCCAGCAGGCG). The gel was visualized using CL-XPosure Film (Thermo Fisher) overnight.

Nitrogen Deprivation and Lipid Analysis

Single colonies from TAP-agar were inoculated into TAP liquid medium and grown under 30-40 μmol photons m^{-2} s^{-1} continuous light at ambient temperature. Cell growth was monitored by measuring the chlorophyll concentration as previously described (Porra et al., 1989). After reaching stationary phase (40-60 μg chlorophyll per mL), the cells were diluted with fresh TAP medium to 1.5 μg/mL chlorophyll and grown for two additional days to reach mid-log phase (5-12 μg/mL chlorophyll). For nitrogen deprivation, cells were pelleted by centrifugation at 1,500 x *g* for 2 min, washed once by resuspending in TAP-N medium and centrifuging at 1,500 x *g* for 2 min, and resuspended with TAP-N medium followed by a 24 h incubation at 30-40 μmol photons m^{-2} s^{-1} continuous light. Cell suspensions containing 30 or 60 μg chlorophyll were collected for lipid extraction.

Lipid extraction and thin-layer chromatography (TLC) were conducted as previously described (Li et al., 2012). For quantitative analysis of TAG, liquid chromatography–mass spectrometry (LC-MS) and tandem mass spectrometry (MS/MS) were performed as previously described (Liu et al., 2013). Glyceryl triheptadecanoate (T2151, Sigma-Aldrich) was added as the internal standard to the lipid extracts. The mixtures were subjected to LC-MS using Agilent 1260 HPLC (Agilent, Santa Clara, CA) and Bruker micrOTOF-Q II mass spectrometer (Bruker, Billerica, MA) as previously reported (Terashima et al., 2014). MS and MS/MS fragmentation data were searched against the LipidMaps database (Sud et al., 2007) to identify specific TAG species. A Mann-Whitney U-test (Mann and Whitney, 1947) was used to assess the significance of the difference between *lcs2* and strains expressing LCS2 (CC-4533 and the complemented lines), using each technical replicate as a separate data point. A false discovery rate correction was performed (Benjamini and Hochberg, 1995).

Molecular Characterization of the Mutants Disrupted in Genes Encoding Lipid Droplet Proteins

To verify the insertions in genes for mutants selected for TLC, PCR was performed as described above for randomly selected mutants (Supplemental Figure 9). Purified genomic DNA obtained by phenol:chloroform:isoamyl alcohol extraction (Zhang et al., 2014) was used as the template. The PCR reaction mix consisted of 5 μL Phusion GC buffer, 0.75 μL DMSO, 0.5 μL 10 mM dNTPs, 0-0.5 μL 50 mM $MgCl_2$, 1.25 μL of each primer, 0.25 μL Phusion HSII polymerase, 25 ng genomic DNA, and sufficient H_2O to total 25 μL. PCR was started with a 30- to 90- s initial denaturation at 98°C, 35 thermal cycles of 15 s at 98°C, 30 s at 63°C, 1 min at 72°C, followed by a final extension of 2 min at 72°C. For RT-PCR for the *LCS2* gene, RNA isolation from log-phase cells and reverse transcription were conducted using the RNeasy Plant Mini Kit (74904, Qiagen) and the RETROscript Kit (AM1710, Ambion, Grand Island, NY) respectively.

Complementation of the *lcs2* Strain

For complementation of the *lcs2* mutant, the bacterial artificial chromosome clone 28N9 (Nguyen et al., 2005) was digested with *Spe*I (New England Biolabs). The 13,292 bp genomic fragment containing the *LCS2* gene (including 5' and 3' UTRs), but no other open reading frame, was gel purified and co-transformed into *lcs2* cells by electroporation, as described above together with *Nde*I-linearized (New England Biolabs) pHyg3 plasmid (Berthold et al., 2002), which confers resistance to hygromycin B). For each cuvette, 115 ng of the 28N9 fragment with 50 ng linearized pHyg3 or 90 ng of the 28N9 fragment with 20 ng linearized pHyg3 were added to the cuvette. A cuvette with no 28N9 fragment and 50 ng linearized pHyg3 was included as a control. Transformants were selected on TAP agar plates containing 15 μg/mL hygromycin B.

Transformants were screened by PCR for the presence of wild-type *LCS2* genomic DNA. The recipe and cycles of PCR are as described above for the validation PCRs for mutants disrupted in genes encoding lipid droplet proteins except that crude DNA extracts were used as templates. To obtain crude DNA extracts, colonies were swirled in 50 μL 5% Chelex-100 (Bio-Rad) in 96-well plates and heated for 12 min at 99°C. Twice during the 99°C incubation, the plates were removed and vortexed for 5 s. The plates were then centrifuged at 2,000 x *g* for 3 min and supernatants transferred to another plate for use as PCR templates. For each 25 μL reaction mix, 2 μL crude extract was used. Approximately 6% of the transformants were positive based on PCR with the primer pair g1 + g2. immunoblots using the peptide antibody described below were used to evaluate the presence of LCS2 protein in the transformants. Four of the 24 PCR positive transformants accumulated detectable levels of the LCS2 protein.

For immunoblotting, log-phase cells were resuspended in 2.4 mM HEPES, 4.8 mM EDTA (pH 7.5) with 0.48 mM benzamidine, 0.48 mM phenylmethylsulfonyl fluoride, 0.48 mM aminocaproic acid, 60 mM dithiothreitol, 60 mM Na_2CO_3 and were solubilized in the presence of 2% sodium dodecyl sulfate and 12% sucrose at 100°C for 55 s. Polypeptides were separated by SDS-PAGE on a 10% SDS-polyacrylamide gel (Bio-Rad). Proteins were transferred onto 0.45 μm PVDF membranes using a semi-dry transfer apparatus (Bio-Rad) at 60 V for 30 min, followed by 150 V for 45 min. Primary antibodies against a peptide epitope of LCS2 (CRRPQLQAKYQAKLD-amide; aa643-656, conjugated with a cysteine and an amide group) were produced in rabbit, affinity-purified by Yenzym (South San Francisco, CA) and used at a 1:1,000 dilution. Secondary goat anti-rabbit antibodies conjugated to horseradish peroxidase (65-6120, Life Technologies), were used at a 1:10,000 dilution to detect the primary antibodies by the chemiluminescence method using the WesternBright ECL Kit (K12045-D20, Advansta, Menlo Park, CA). For detection of the tubulin loading control, the PVDF membrane was stripped using RestorePLUS Western Blot Stripping Buffer (PI46430, Thermo Fisher Scientific), and reblotted using mouse anti-α-tubulin primary antibodies (T6074, Sigma-Aldrich) in a 1:10,000 dilution, followed by goat anti-mouse secondary antibodies conjugated to horseradish peroxidase (G-21040, Life Technologies) in a 1:10,000 dilution, and were then detected by chemiluminescence using the Amersham ECL Western Blotting Detection Reagents (RPN2106, GE Healthcare).

Additional Software Used for the Analysis

Several python packages were used at various stages in the analysis: NumPy (Oliphant, 2007) was used throughout the code for large array and numerical operations; matplotlib (Hunter, 2007) was used for visualization; SciPy (http://www.scipy.org) and RPy2 (http://rpy.sourceforge.net) were used for statistical functions; Biopython (Cock et al., 2009) was used for sequence file parsing.

SUPPLEMENTAL REFERENCES

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B **57,** 289-300.

Berthold, P., Schmitt, R., and Mages, W. (2002). An engineered Streptomyces hygroscopicus aph 7" gene mediates dominant resistance against hygromycin B in Chlamydomonas reinhardtii. Protist **153,** 401-412.

Betsumiya, K., Gulliver, T.A., and Harada, M. (2003). Extremal Self-Dual Codes over F2 x F2. Designs, Codes and Cryptography **28,** 171-186.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) **6,** 80-92.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25,** 1422-1423.

Collins, S.R., Schuldiner, M., Krogan, N.J., and Weissman, J.S. (2006). A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biol **7,** R63.

Crutchfield, A.M., Diller, K.R., and Brand, J.J. (1999). Cryopreservation of Chlamydomonas reinhardtii (Chlorophyta). Eur. J. Phycol. **34,** 43-52.

Gallaher, S.D., Fitz-Gibbon, S.T., Glaesener, A.G., Pellegrini, M., and Merchant, S.S. (2015). Chlamydomonas Genome Resource for

Laboratory Strains Reveals a Mosaic of Sequence Variation, Identifies True Strain Histories, and Enables Strain-Specific Studies. Plant Cell, [Epub ahead of print].

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Computing In Science & Engineering **9,** 90-95.

Kropat, J., Hong-Hermesdorf, A., Casero, D., Ent, P., Castruita, M., Pellegrini, M., Merchant, S.S., and Malasarn, D. (2011). A revised mineral nutrient supplement increases biomass and growth rate in Chlamydomonas reinhardtii. Plant J **66,** 770-780.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10,** R25.

Li, H. (2011). Improving SNP discovery by base alignment quality. Bioinformatics **27,** 1157-1158.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25,** 1754-1760.

Li, X., Benning, C., and Kuo, M.H. (2012). Rapid triacylglycerol turnover in Chlamydomonas reinhardtii requires a lipase with broad substrate specificity. Eukaryot Cell **11,** 1451-1462.

Liu, B., Vieler, A., Li, C., Jones, A.D., and Benning, C. (2013). Triacylglycerol profiling of microalgae Chlamydomonas reinhardtii and Nannochloropsis oceanica. Bioresour Technol **146,** 310-316.

Mann, H.B., and Whitney, D.R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Annals of Mathematical Statistics **18,** 50-60.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal **17,** 10-12.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20,** 1297-1303.

Nguyen, R.L., Tam, L.W., and Lefebvre, P.A. (2005). The LF1 gene of Chlamydomonas reinhardtii encodes a novel protein required for flagellar length control. Genetics **169,** 1415-1424.

Oliphant, T.E. (2007). Python for Scientific Computing. Computing In Science & Engineering **9,** 10-20.

Porra, R.J., Thompson, W.A., and Kriedemann, P.E. (1989). Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls a and b extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. Biochimica et Biophysica Acta (BBA) - Bioenergetics **975,** 384-394.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol **29,** 24-26.

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E.A., Glass, C.K., Merrill, A.H., Jr., Murphy, R.C., Raetz, C.R., Russell, D.W., and Subramaniam, S. (2007). LMSD: LIPID MAPS structure database. Nucleic Acids Res **35,** D527-532.

Terashima, M., Freeman, E.S., Jinkerson, R.E., and Jonikas, M.C. (2014). A fluorescence-activated cell sorting-based strategy for rapid isolation of high-lipid Chlamydomonas mutants. Plant J **81,** 147-159.

Zhang, R., Patena, W., Armbruster, U., Gang, S.S., Blum, S.R., and Jonikas, M.C. (2014). High-Throughput Genotyping of Green Algal Mutants Reveals Random Distribution of Mutagenic Insertion Sites and Endonucleolytic Cleavage of Transforming DNA. Plant Cell **26,** 1398-1409.